

# 马达数据后端系统项目

2016年4月25日

## 注意

关于本项目的信息和要求请你严格保密，请勿将本项目要求信息公开到网上。你自己的项目代码和文档本身可以放到你自己的 Github repo 上，但是请尽量避免出现关于马达数据公司和面试流程的信息。

## 项目介绍与要求

### 数据获取

[Crunchbase](#) 是创业圈非常有名的媒体和数据库，他们搜集和整理了各个创业公司、投资人、创业者和公司上市的信息。

Crunchbase 已经把2013年的数据整理公开，放到了这个链接：<https://data.crunchbase.com/docs/2013-snapshot>。

数据库中包含以下几个表：

+-----+	
	Tables_in_crunchbase
+-----+	
	cb_acquisitions
	cb_degrees
	cb_funding_rounds
	cb_funds
	cb_investments
	cb_ipos
	cb_milestones
	cb_objects
	cb_offices
	cb_people
	cb_relationships
+-----+	

大致包含了公司的信息、投资机构的信息、上市事件以及一些关键人物的信息及其关系网。

你可以通过注册申请获得数据。如果在这一步卡住了，请随时联系我们帮忙。

数据本身是通过mysqldump出来的 SQL 文件，你可以直接导入到一个新的mysql数据库中。

## 项目要求

有了以上的数据，我们就可以做很多有意思的事情了（比如，其中的cb\_people表就包含了226709个数据）。

你可以用你钟爱的 ORM 工具（比如 [JDBI](#)、[JOOQ](#)、[SQLAlchemy](#)、[Sequelize](#)、[Ecto](#)）来访问这些数据，在你钟爱的语言里面建模、处理这些数据。你还可以用你最喜欢的后端框架（比如 [Dropwizard](#)、[Flask](#)、[Express](#)、[Phoenix](#)）来提供一系列 Restful API，进一步方便前端 JavaScript 可视化。

为了让这个项目本身更有意思一些，我们希望你来试一试一个比 Restful 更灵活的 API 框架 [GraphQL](#)，并且以此为基础，结合你最喜欢的后端框架，搭建一个灵活的探索 Crunchbase 数据集的 API 服务，让用户可以方便的问（即：写 GraphQL）例如以下几个问题：

- 所有融资过 B 轮的公司，他们的团队和其毕业学校是哪些？
- 有哪些是斯坦福背景的投资人投资了斯坦福背景的公司？
- 投资过 Facebook 的所有投资人他们所属的投资机构（或者个人）还投资过哪些公司？

请注意我们的要求并不是你只实现以上几个 API，而是希望你以他们为例子，结合你学到的 GraphQL 的知识，在你喜欢的编程语言里面搭建合适的模型，定义合适的 Query 和 Schema Type 以及 Input Params，来支持更广泛、通用和深入的查询。比如把最后一个问题中的 Facebook 换成 Google，或者把投资过哪些公司变成毕业于哪些学校，或者更深入地问这些公司的投资人又是谁，还可以更深入的问这些投资人又投资过哪些公司。GraphQL让你能用简单的代码灵活地回答以上的问题，而不是根据特定的问题单独写出 function 来对应。

比如，一个 query 的例子可以是：

```
query {
  objects(where: { name: "Palantir" }) {
    ... on Company {
      name
      domain
      homepageUrl
      overview
      foundedAt
      allProducts {
        name
      }
      fundingRounds
      allFundingRounds {
        fundingRoundType
        fundedAt
        raisedAmountUsd
        preMoneyValuationUsd
        postMoneyValuationUsd
        sourceDescription
        allInvestors {
          name
          overview
          investedCompanies
          allRelationships {
            title
            startAt
            relationshipTarget {
              name
              overview
            }
          }
        }
      }
    }
  }
}
```

```
}  
}
```

通过这个 query 我们可以知道所有名字里面包含 Palantir 的公司的所有产品名字、所有融资轮次及其投资人，以及这些投资人有关联的公司的名字。

P.S. 以上几个问题，做完之后你再比较一下如果用传统的 Restful API 来写的复杂度和用 GraphQL 的复杂度。

## 我们建议的解决问题的思路

- 请首先研究清楚数据库中表格之间的关连（which fields in which tables reference other tables?）
- 选择一个合适的、你熟悉的 ORM 工具，并且利用好 ORM 里面的一对多关系、多对多关系映射
- 利用好 GraphQL 对于关系型查询的优势，而不是自己去写业务逻辑代码实现上述功能（再次强调一下，我们的要求并不是只是实现上述几个查询，而是设计一个 GraphQL API 可以让用户自由的实现多种关系型查询）

## Tips和注意事项

我们的项目考察的目的是：

- 沟通交流能力
- 你在短时间内对新工具的学习领悟能力
- 数据建模和编码的能力
- 代码风格和编程习惯

你也可以参考以下一些资源：

- <https://github.com/chentsulin/awesome-graphql> 一个社区整理的 GraphQL 本身的资源，其中的 GraphiQL 特别值得推荐
- <https://github.com/akullpp/awesome-java#orm> 一个社区整理的关于 Java 和 Java ORM 的资源列表，关于其他编程语言，相信你也可以 Google 到